A COMPARISON OF CLUSTERING METHODS AFFECTION ON GENETIC REGULATORY NETWORKS

F. Bakouie[1], M. H. Moradi[2]
[1,2]Biomedical signal processing laboratory, Faculty of Biomedical Engineering ,Amirkabir University of Technology (Tehran Polytechnich), Tehran 15875 4413, Iran

e-mail: fbakouie@cic.aut.ac.ir
mhmoradi@aut.ac.ir

Abstract- **Gene regulatory networks model regulation in living organism. For constructing some of these networks we require prior biological knowledge. Clustering algorithms are used for investigating this information via considering dependence and independence in gene microarray data. There are lots of similarity criteria in clustering methods that each of them can affects on clustering results. So performance of clustering algorithms is important in modeling genetic regulatory networks. We compare affection of three clustering methods on a proposed genetic regulatory network. These methods are: clustering based on mutual information similarity and simulated annealing optimization, Hierarchical clustering based on mutual information distance, and Variable Neighborhood Search (VNS) clustering. For evaluating these clustering methods on regulatory networks, we used a simple nonlinear model as our genetic regulatory network and a new algorithm for calculating parameters of model by using Least Square (LS) estimation. For evaluating our model, Coefficient of Determination (CoD) is used.**

Keywords **–Gene microarray, genetic regulatory network, clustering, mutual information, Coefficient of Determination (CoD), Least Square (LS) estimation.**

I. INTRODUCTION

Traditional molecular biology typically focuses on a single gene, protein, reaction or pathway, and follows a reductionist approach to studying the biological system. Over the years, this practice has led to remarkable achievements. However, biological processes are inherently integrated and interactive, so traditional studies cannot resolve the complex relationships among biological entities. Therefore to understand the nature of cellular function, it is necessary to study the behavior of genes in a holistic rather than in an individual manner because the expressions and activities of genes are not isolated or independent of each other [1].

The major challenge in the post-genome era is to understand how interactions among genes in a cell determine its form and function. So we need to develop methodologies for identifying and analyzing the complex biological networks that regulate metabolism. Regulatory networks modulate the action of metabolic networks, leading to physiological and morphological changes. Even though new high-throughput transcriptomic, proteomic, and metabolomic analysis technologies give biologists vast amounts of valuable data, techniques that model uncertainty are needed to cope with the many genes of uncertain function and to understand complex interactions[2].

In related studies variety of mathematical and computational methods are being developed in order to construct formal models of genetic interactions. There have been a number of attempts to model gene regulatory networks, including linear models, Bayesian networks, neural networks, differential equations, and models including stochastic components on the molecular level [1]. Boolean networks (BNs) are proposed to provide insight and a conceptual framework for an integrative view of genetic function and regulation [3]. How to build BNs from gene microarray data remains an open problem. The advantage of BNs is that they naturally allow one to incorporate prior biological knowledge [3]. Thus, if certain regulatory relationships are known to exist, then the Boolean-function classes corresponding to the genes in question can be constructed to reflect this prior knowledge.

One useful way for investigating the prior biological knowledge is via an unsupervised clustering algorithm by considering dependence and independence in gene microarray data. Clustering has been used in a number of studies to obtain a global, unsupervised perspective on the similarity of expression problems [1,4,5,6]. Each of these methods considers a special criterion as similarity among genes that are in the same cluster. In this paper we evaluate affection of three clustering algorithms on merit of genetic regulatory network which is constructed based on knowledge is obtained from clustering algorithms that are: clustering based on mutual information similarity and simulated annealing optimization, Hierarchical clustering based on mutual information distance, and Variable Neighborhood Search (VNS) clustering.

This paper is organized as follows. Section II describes the problem of constructing genetic regulatory networks; our approach in modeling, Coefficient of Determination (CoD) for evaluating our model and different clustering methods which is used in this paper. Section III provides experimental results. In section IV we discuss about the affection of clustering methods on merit of genetic regulatory networks which is constructed based on them. Finally in section V we conclude and propose new approach for investigating prior biological knowledge.

II. METHODOLOGY

A. Gene Regulatory Nnetworks

A gene regulatory network $G(V,F)$ consists of a set $V=\{x_1,\ldots,x_n\}$ of nodes representing genes and a list

$F = (f_1, \ldots, f_n)$ of functions, where a function $f_i(x_{i_1}, \ldots, x_{i_k})$ with inputs from specified nodes $x_{i_1}, \ldots, x_{i_k}$ is assigned to each node x_i. The function f_i are also referred to as predictor of gene i [1]. In order to build the network for gene x_i, we implement the following four steps:

(1) Determine the input set of variables corresponding to function f_i. This is done by using a clustering technique. In this paper we use three methods. We will describe each of them later.

(2) Each f_i is then modeled by a nonlinear model called perceptron [7]. For predicting a target expression value *Y* (for example Y is gene i : x_i) from predictor variables $X_1, X_2, ..., X_m$, a perceptron takes the form of

$$Y_{pred} = g(a_1 X_1 + a_2 X_2 + ... + a_m X_m + b), \qquad (1)$$

where *g* is a threshold function [7]. Design of a perceptron requires estimating the coefficients $a_1, \ldots, a_m$ and b. In [7] these parameters were calculated by using a training algorithm while in this paper we use Least Square (LS) estimation.

(3) The coefficient of determination (CoD) is employed to compute merit of our prediction. The determination coefficient is defined in accordance with the degree to which a filter estimates a target variable beyond the degree to which the target variable is estimated by its mean [7,8]. So the bigger CoD shows better filtering and prediction.

B. Clustering Based on Mutual Information and Simulated Annealing Optimization

The motivation for considering mutual information is its capacity to measure a general dependence among random variables. The main idea of this technique is to identify a relatively small number of candidate parents for each gene based on statistics such as correlation. We then restrict our search to networks in which only the candidate parents of a variable can be its parents, resulting in a smaller search space in which we can more efficiently find a good structure.

Shannon's information theory provides a suitable formalism for quantifying the above concepts. Suppose we are to partition the set of gene variables into k disjoint subsets as $V = X_1 \cup X_2 \cup \ldots \cup X_k$. The cost function is defined as the sum of pair-wise mutual information between any two subsets,

$$f(s) = \sum_{i \neq j} I(X_i; X_j), \qquad (2)$$

where s denotes a particular partition scheme and $I(X_i; X_j)$ is mutual information between cluster i and j. The mutual information between X and Y is a measure of information about X (or Y) contained in Y (or X) and given by

$$I(X;Y) \equiv H(X) - H(X|Y), \qquad (3)$$

Where H(X) is entropy of X [1,4].

The simulated annealing (SA) algorithm is employed to find an optimal partition scheme such that the cost function attains the minimum [1, 9].

C. Hierarchical Clustering Based on Mutual Information Distance

In this method we implement the following scheme for clustering n objects (genes) in to k clusters.

(1) Compute a proximity matrix based on pair wise mutual information distance; assign n clusters such that each cluster contains exactly one object.
(2) Find the two closest clusters i and j;
(3) Create a new cluster (ij) by combining i and j;
(4) Delete the lines/columns with indices i and j from the proximity matrix, and add one line/column containing the proximities between cluster (ij) and all other clusters;
(5) If the number of clusters is still > k, go to (2); else join the two clusters and stop [4].

Our proximity scale is based on relative distance which is defined via mutual information [4]

$$D(X,Y) = 1 - \frac{I(X,Y)}{H(X,Y)}, \qquad (3)$$

C. Variable Neighborhood Search (VNS) Clustering

The VNS algorithm is a recently proposed metaheuristic for solving combinatorial and global optimization problems. The basic goal of the method is to proceed to a systematic change of neighborhood within a local search algorithm. This algorithm remains in the same locally optimal solution exploring increasingly far neighborhoods by random generation, until another solution better than the incumbent is found. When so, it jumps to the new solution and proceeds from there. The neighborhood centroid structures are obtained by replacing at random some predetermined number *k* of existing centroids of clusters with *k* randomly chosen patterns, i.e., genes. For a more detailed analysis of VNS metaheuristic, see references [6, 10]. In this paper we use General VNS (GVNS) method for clustering [6].

III. RESULTS

Our experiments are based on the observations in transcription level in the context of responsiveness to genotoxic stresses. The ternary data of the survey (14 genes and 30 samples) are given in [1, 7]. We model regulatory network just for 4 genes: x12, x11, x10 and x2. For modeling a network, at first we should find parent genes for each target gene. It is done via clustering results. As mentioned in [1] we have done clustering for different number of clusters. In this work the number of clusters varies from 2 to 7, so we have different set of parent genes for each target gene and we can model network in different level of dependences. Table I, II

and III show results of networking for each method of clustering.

TABLE I
MODELING BASED ON MI CLUSTERING and SA OPTIMIZATION

Parent genes for gene x12	*CoD*
(x1,x3,x4,x5,x10,x11,x13,x14)	0.654
(x1,x4,x7,x13)	0.653
(x5,x8,x10)	0.203
(x6,x10,x14)	0.254
(x3)	0
(x14)	0

Parent genes for gene x11	*CoD*
(x1,x3,x4,x5,x10,x12,x13,x14)	0.073
(x3,x5,x10,x14)	0.134
(x1,x3,x5)	0.016
(x4,x13)	0.0325
(x2)	0.13

Parent genes for gene x10	*CoD*
(x1,x3,x4,x5,x11,x12,x13,x14)	0.198
(x3,x5,x11,x14)	0.263
(x5,x8,x12)	0.290
(x6,x12,x14)	0.306
(x6,x14)	0.184
(x7)	0

Parent genes for gene x2	*CoD*
(x6,x7,x8,x9)	0.315
(x6,x8,x9)	0.316
(x4,x13)	0.197
(x8)	0.07
(x11)	0.072
(x14)	0.069

TABLE II
MODELING BASED ON VNS CLUSTERING

Parent genes for gene x12	*CoD*
(x1,x13)	0.749
(x13)	0.578

Parent genes for gene x11	*CoD*
(x2,x3,x4,x5,x6,x7,x8,x9,x10,x14)	0.033
(x5,x14)	0.013
(x5)	0.009

Parent genes for gene x10	*CoD*
(x2,x3,x4,x5,x6,x7,x8,x9,x11,x14)	0.108
(x1,x2,x3,x4,x6,x7,x8,x9)	0.186
(x2,x3,x4,x6,x7,x8,x9,x14)	0.173
(x2,x3,x6,x8,x9)	0.327
(x3,x14)	0.190

Parent genes for gene x2	*CoD*
(x3,x4,x5,x6,x7,x8,x9,x10,x11,x14)	0.143
(x1,x3,x4,x6,x7,x8,x9,x10)	0.270
(x3,x4,x6,x7,x8,x9,x10,x14)	0.297
(x3,x6,x8,x9,x10)	0.241
(x4,x6,,x7,x8,x9)	0.256

IV. DISCUSSION

As we see from tables I, II and III different clustering methods propose different parent genes for each target gene, the question is what kind of clustering gets precise information for making regulatory networks. In this paper we use three different clustering methods, we use hierarchical and non hierarchical clustering scheme. Although our statistic criterion is different, we use mutual information that considers some nonlinear dependences and VNS that use k-means as its local search. But we can't say what the best method is. For example consider genex12, by using VNS clustering results as prior knowledge we have a the best modeling for this target gene whit CoD=0.749 while for gene x11 clustering based on MI clustering with SA optimization gets the best clustering with CoD=0.136. For gene x2 hierarchical clustering leading to the best modeling with CoD=0.316.

V. CONCLUSION

By considering results of this work it seems that we should have a revision on our clustering methods and their criteria as dependences and similarity among genes in genetic regulatory networks. In existing clustering techniques we model biological similarity with statistical criteria but it seems that stistics by itself can't model biological dependences well, so, for obtaining prior biological knowledge by using just statistical algorithms we won't have a good model as regulatory networks.

TABLE III
MODELING BASED ON HIERARCHICAL CLUSTERING

Parent genes for gene x12	*CoD*
(x1,x6,x8,x9,x10,x11,x13,x14)	0.627
(x1,x6,x8,x9,x13)	0.649
(x13)	0.58

Parent genes for gene x11	*CoD*
(x1,x2,x4,x6,x7,x8,x9,x10,x12,x13)	0.0262
(x1,x6,x8,x9,x10,x12,x13,x14)	0.136
(x10,x14)	0.017
(x10)	0.055

Parent genes for gene x10	*CoD*
(x1,x6,x8,x9,x11,x12,x13,x14)	0.121
(x11,x14)	0
(x11)	0

Parent genes for gene x2	*CoD*
(x1,x4,x6,x8,x9,x10,x11,x12,x13)	0.223
(x4,x7)	0.135

ACKNOWLEDGMENT

This work is supported by Iran Telecommunication Research Center (ITRC).

REFERENCES

[1] X. Zhoua, X Wangb, E. Doughertya; "Construction of genomic networks using mutual-Information clustering and reversible-jump Markov-chain-Monte-Carlo predictor design," Signal Processing, vol. 83, pp. 745 – 761, 2003

[2] P. Du, J. Gong, E.S. Wurtele, and J.A. Dickerson, "Modeling Gene Expression Networks Using Fuzzy Logic" IEEE transactions on system,man,and cybernetics-partB:cybernetics,vol. 35, no.6, pp 1351-1359, december 2005 .

[3] I. Shwulevich, E.R. Dougherty, S. Kim and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," Bioinformatics, vol.18. no.2, pp. 261-274, 2002

[4] A. Kraskov, H. St¨ogbauer, R. G. Andrzejak, and P. Grassberger, " Hierarchical Clustering Using Mutual Information ",2003

[5] X, Zhou, X. Wang, E.R. Dougherty, D. Russ, and E. Suh, " Gene Clustering Based on Clusterwide Mutual Information," Journal of computational biology, Volume 11, © Mary Ann Liebert, Inc. Pp. 147–161, Number 1, 2004

[6] F. Bakouie, MH. Moradi, " application of Variable Neighborhood Search (VNS) methods in microarray data clustering," ICEE2006, 14th Conference on Electrical Engineering, Iran, Tehran, Amirkabir University of technology.

[7] S. Kim, E. R. Dougherty, M. L. Bittner, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, "General nonlinear framework for the analysis of gene interaction via multivariate expression arrays," Journal of Biomedical Optics, vol.5, pp. 411–424, October 2000

[8] E. R. Dougherty, S. Kim, Y. Chen, "Coefficient of determination in nonlinear signal processing," Signal Processing, vol. 80, pp. 2219-2235, 2000

[9] L. Ingber, B. Rosen, "genetic algorithm and very fast simulated reannealing: a comparison," Mathematical and Computer Modelling, vol. 16(11), pp. 87-100, 1992

[10] N. Belacel, M. Cuperlovic-Culf, R. Ouellette, and M. Boulassel, "The Variable Neighborhood Search Metaheuristic for Fuzzy Clustering cDNA Microarray Gene Expression Data," Proceedings of IASTED-AIA-04 Conference. Innsbruck, Austria, February 16-18, 2004